WHAT IS CLAIMED IS:

1. A cache memory for three-dimensional graphics texture mapping, comprising:

first and second DRAM banks including SAM ports, respectively, each of said SRAM ports reading a texture for a trilinear interpolation and fetching new texture sub-clips from the outside;

a sub-clip loader connected to said SAM ports of said first and second DRAM banks and for fetching new texture sub-clips from an external memory;

a controller for controlling said components; and

a CAM for checking if eight texels existing at an integer coordinate relative to an LOD and (u, v) coordinates are located in said first and second DRAM banks, when the LOD and (u, v) coordinates mapped into a texture space with respect to a pixel to be rendered on a display screen are input to said controller.

2. The cache memory according to Claim 1, wherein said first DRAM bank forms a clip RAM pyramid for storing all texels associated with several upper levels of the LOD among overall mipmaps, and wherein said second DRAM bank forms a clip RAM stack for storing only a working set

currently needed among the remaining LOD levels being not stored in said first DRAM bank.

3. The cache memory according to Claim 1, wherein said second DRAM bank has a plurality of n x n sub-clips obtained by dividing a texture area on one level of LOD of the mipmap, and the contents of said cache memory are replaced on the sub-clip basis.

4. The cache memory according to Claim 1, wherein data paths are provided between said first and second DRAM banks and a trilinear interpolator, said data paths being adapted to access eight texels existing at the integer coordinate locations centered on (u, v) coordinates and LOD mapped into a texture space of a pixel to be rendered on a display screen, such that a trilinear interpolation is performed in one clock cycle.

5. The cache memory according to ~~any one of Claims 1 to 4,~~ Claim 1 further comprising a sub-clip predictor for performing a hardware-based prefetch of a sub-clip to be soon needed, so as to reduce a penalty due to cache miss.

6. A method for reducing a penalty occurring upon a

cache miss, comprising the steps of:

performing a sub-clip prediction in one stack layer, where, under a hardware-based sub-clip prediction limit of 2 by 2 sub-clip boundary inside of sub-clips (4 by 4) on a current clip RAM stack, when the tracing of (u, v) coordinates passes the limit, sub-clips of the tracing direction (left, right, upper and lower sides) are prefetched; and

performing a stack layer prediction where the current clip RAM stack represents the levels of LOD (LOD i to LOD i+3) stored in said stack, internal two levels of LOD are used as a prediction limit, and, immediately after the tracing of LOD passes the limit, a stack layer corresponding to a next level of LOD is prefetched.